Journal of Biomedical Informatics 44 (2011) 997–1003

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Computer-based genealogy reconstruction in founder populations

Giuseppe Milani [b,*,1], Corrado Masciullo [a,1], Cinzia Sala [a], Riccardo Bellazzi [b], Iwan Buetti [a], Giorgio Pistis [a], Michela Traglia [a], Daniela Toniolo [a,c,1], Cristiana Larizza [b,1]

[a] Division of Genetics and Cell Biology, San Raffaele Scientific Institute, 20132 Milano, Italy
[b] Department of Computer Engineering and Systems Science, University of Pavia, 27100 Pavia, Italy
[c] Institute of Molecular Genetics – CNR, 27100 Pavia, Italy

ARTICLE INFO

Article history:
Received 28 May 2010
Accepted 4 August 2011
Available online 23 August 2011

Keywords:
Population genetics
Data integration
Algorithms
Record Linkage
Pedigree

ABSTRACT

This paper describes a software tool that reconstructs entire genealogies from data collected from different and heterogeneous sources, including municipal and parish records archived over centuries. The tool exploits a record linkage algorithm relying on a rule-based data matching approach. It applies a general strategy for managing the ambiguities due to missing, imprecise or erroneous input data. The process follows an iterative approach that combines automatic pedigree reconstruction with software-empowered human data revision to improve the quality and the accuracy of the results and to optimize the matching rules.

The paper discusses the results obtained by reconstructing the entire genealogy of the population of the Val Borbera, a geographically isolated valley in Northern Italy. The genealogy could be reconstructed from data going back as far as the XVI century. The resulting pedigree includes 75,994 trios, 58.9% of which belonging to a unique big family, reconstructed over 13 generations.

1. Introduction

Genetic analysis in large families is a very powerful approach in genetic studies and has resulted in many successful strategies to identify thousands of genetic loci and genes for Mendelian and complex disorders [1]. An alternative to "traditional" families or to sib-pairs analysis is the use of extended families, particularly of large genetically isolated or inbred populations [2]. Such populations have been successfully used for the study of many rare Mendelian disorders. Due to their derivation from few common ancestors, to their decreased genetic heterogeneity and the common environment, it was proposed that these populations might be also suitable for identification of genetic risk factors for common disorders [3]. In many cases, this was indeed the case [4–6].

To completely reconstruct the genealogy and determine the distant relationships between today descendants from the common ancestors is not a simple task. Family relationships over many generations are rarely available in electronic format in isolated populations: one such case was the Older Amish Population from Pennsylvania [7]. For relatively small populations, in the range of few to hundred thousands inhabitants as a maximum size, a complete genealogy that includes all living population and most of their ancestors can however be reconstructed thanks to the availability of church records going back for centuries and conserved in church archives. Several approaches have been used to reconstruct population genealogies, depending from the size and the structure of the population. In most instances partial family pedigrees were available and the whole genealogy has been semi-manually reconstructed and subsequently computerized [8,9]. In most projects, and one example is the BALSAC Project [http://www.uqac.ca/balsac/ang/index.php], a large manual effort was required to reconstruct the pedigree. Within BALSAC, the Quebec population pedigree was reconstructed from the database of the population vital records starting from marriage acts. The record linkage is based on perfect matching of the identification data (name and surname) of the couples reported. Every time the match is not perfect an expert tries to reconstruct the families, manually.

The main problem for automatic reconstruction of the complete genealogy of a population is the variety of sources to be considered that contain information necessary to uniquely link each individual to one's parents and the many inconsistencies across multiple entries. Moreover, many people have the same first and last names and their birth dates are often quite imprecise.

Several tools can be used today for pedigree reconstruction from genotype data, now available for many populations. Examples are FRANz, a software for pedigree reconstruction in natural populations using co-dominant genomic markers [10], KINGROUP, a program for pedigree relationship reconstruction and kin group assignments using genetic markers [11] and KINALYZER, [12].

* Corresponding author. Address: Department of Computer Engineering and Systems Science, University of Pavia, Via Ferrata, 1, 27100 Pavia, Italy.
E-mail address: milani.giuseppe@gmail.com (G. Milani).
[1] These authors equally contributed to the paper.

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.08.004

Finally, many large databases exist offering access to genealogical information that may provide small pedigree reconstruction (see: http://www.kindredtrails.com/databases.html or http://www.myheritage.com/) but to our knowledge, a tool for the visual representation of a manually reconstructed pedigree of an entire population was reported only once [13]. We present here a tool that can be used to construct nearly automatically a genealogy of a medium size population from different sources and from data collected over many centuries. The tool can perform in a single run every step from data preparation to pedigree generation and was tested in the reconstruction of the genealogy of the population of Val Borbera, a geographically isolated valley in North West Italy. Using the tool it was possible to reconstruct the entire pedigree and link most of today population in a large genealogical tree going back to the population living in the valley from the middle of 1500.

2. Materials and methods

The whole pedigree reconstruction process is based on an iterative procedure (we will refer to every iteration as a "run") based on four main steps (Fig. 1): the first two steps (data import/cleaning) automatically integrate data from different repositories into a database suitable for pedigree reconstruction and run some data preprocessing procedures. The third step generates the pedigree and several reports relating the results of the current run and describing the ambiguous or inconsistent cases found in the database. The archivists manually carry out the fourth step: they analyze the reports in order to detect possible data inconsistencies preventing the complete reconstruction of the pedigree. Manual data revision is also necessary to tune or optimize the data matching rules used for the pedigree reconstruction, to deal with exceptions, like handling name and surname variations due to the language changes over the centuries, as will be explained in Section 2.2.1.

Each run outputs the population pedigree as a collection of trios (child–father–mother) of those individuals whose information is precise enough to derive their parents. The data revision step can be repeated anytime new data are added to the electronic archives. As the data collection is a long and laborious task, the pedigree reconstruction can be performed progressively over any time span. After each data revision step the automatic reconstruction is re-executed to obtain a new pedigree version.

In the following we will describe the four main steps reported in Fig. 1: data import, data cleaning, pedigree reconstruction and reports evaluation.

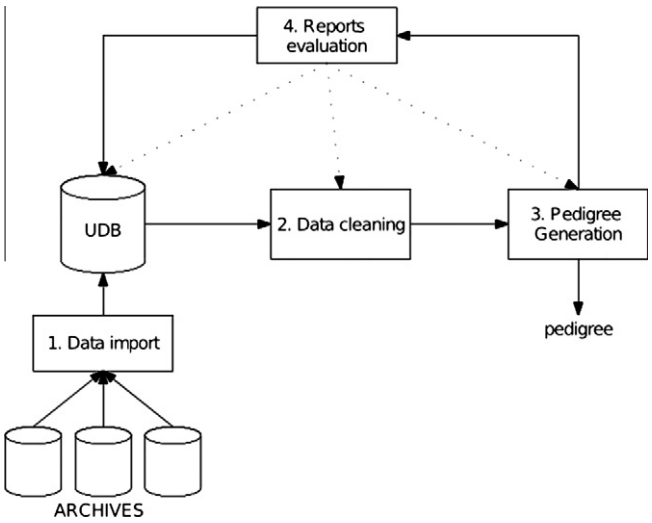

Fig. 1. The overall architecture of the pedigree reconstruction algorithm.

2.1. Data import

As mentioned above, the algorithm exploits data sources that contain the different kind of information necessary to reconstruct the pedigree without ambiguities. Data coming from official vital records are integrated with genealogical information available from other records collected directly from the population or from any other available source. Official vital records include birth, death and marriage registries, on paper or in electronic format (indicated as EMR, Electronic Municipal Records). A thorough evaluation of the records of the Val Borbera population showed some characteristics common to all records:

- Every record contains demographic data on several individuals and the relationships between them.
- The individuals reported in each record are the record owner(s) and his/her/their relatives. The owner might be one of: child (birth record), dead (death record), spouses (marriage record). The relatives can be parents or grandparents.
- The detected relationships are: mother, father, grandfather, grandmother, husband, wife.

Given a set of data sources, the import procedure first extracts the individuals from each archive, keeping track of their relationships and of the data source; then, after proper standardization and data preparation, it stores the data into a Unique Data Base (UDB) structure independent from the data source. Individuals and relationships are put into two distinct tables: the INDIV table, which contains a record for each individual found in the source databases and the RELAT table, which reports the relationship between couple of individuals (e.g. father, mother, grand-father, grand-mother, etc.) [14].

Each individual is classified according to his/her role in the source record. The 1st category group includes the owners of birth and death records. The 2nd category is assigned to the owners of marriage records (the two spouses) and, finally, the 3rd category corresponds to any other individual reported in the records and not belonging to the previous categories. Regardless of the category, all individuals are used during the pedigree reconstruction as they provide information on the family relationships. In particular, 3rd category individuals are used to link to each other the individuals belonging to the 1st category, while 2nd category individuals are mandatory to confirm a child-parent relationship.

As an example, let us consider the import of a birth record into the UDB. This process generates three tuples, recorded into the INDIV table, which correspond to the individuals reported in the birth act (owner, father, mother), and two tuples, which are recorded into the RELAT table to store the two relationships *father* and *mother*.

During the data preparation process the algorithm tries also to complete the missing/inaccurate personal data from the available information [15]. This step is necessary for deriving a set of attributes serving as "key" for linking records that could refer to the same individual. In particular, specific rules have been implemented to compute missing birth and death dates and to impute surnames or places of birth derived from other information.

For example, we defined rules for deriving:

- birth date from age or death date,
- birth/death date from the birth/death record registration date,
- place of birth/residence from the record registration place,
- father/paternal grandfather's surname from the son's/grandson's surname,
- death date from the birth date (and vice versa),
- parents' birth date from the son's birth date.

Since it may frequently happen that only the approximate age of a person is available, every date is represented in the UDB as an

interval. Let's denote with DoB the true Date of Birth and with $\Delta DoB = [DoB_{from}, DoB_{to}]$ a time interval which contains date of birth. As an example, the rule for deriving the father's birth date interval ($father.\Delta DoB = [father.DoB_{from}, father.DoB_{to}]$) from the son's birth date ($son.\Delta DoB = [son.DoB_{from}, son.DoB_{to}]$) is the following:

$$[father.DoB_{from}, father.DoB_{to}] = [(son.DoB_{from} - PA_{max}), (son.DoB_{to} - PA_{min})]$$

where PA_{min} and PA_{max} are the (assumed) minimal and maximal procreative age (PA) of a men.

2.2. Data cleaning

Before running the pedigree reconstruction algorithm, a two-phase data cleaning step has to be performed: the first phase standardizes individual names and surnames, the second eliminates duplicates from the INDIV table as described in the following sections.

2.2.1. Name and surname standardization

As the data come from heterogeneous and often handwritten sources, names, surnames and places are often differently spelled or contain spelling errors and abbreviations heavily lowering the performance of the pedigree reconstruction algorithm.

To perform name and surname standardization we resorted to the following strategies:

- *Automated names normalization* names whose distance is lower than a predefined threshold are clustered together. The algorithm uses a similarity metric called Damerau–Levenshtein distance (DL) [16,17], which is a generalization of the more popular Levenshtein distance [18]. In order to handle the cases in which names contain more tokens, we have implemented a suitable adaptation of the DL distance.
- *Manual names normalization* archivists build a look-up table containing all exceptions (in terms of pair of names) that need to be specified to cover cases not properly handled by the similarity metric (for example "Giovanni Battista" and "Giobatta" correspond to the same name, although their DL distance is quite high).
- *Manual clustering of surnames* archivists manually cluster surnames taking into account the evolution of the language over historical periods. For example, in VB the current surname "Cogo" was reported as "De Cuoghis" or "De Coghi" before the XX century. All variations of each surname were clustered under the modern name.

2.2.2. Duplicates elimination

This step is necessary to delete duplicated 1st and 2nd category individuals, in order to avoid the generation of multiple relations. As a matter of fact, the same 1st category individual can be reported in multiple records (e.g. in a birth and in a death record), which can contain identical or slightly different identification data, while 2nd category individuals can have multiple instances, as marriage records were frequently registered in the birth places of both spouses.

Many approaches can be adopted for record linkage. They can be classified as probabilistic approaches, like the one proposed by Newcombe [19,20] and formalized by Fellegi and Sunter [21], and methods that rely on domain knowledge or on generic distance metric [22,23]. More recently, several supervised classification methods have been also proposed [24,25].

The duplicate detection task in our system is based on the equational theory approach proposed by Hernandez and Stolfo [26]. This theory, called rule-based approach, specifies an inference about the similarity of records and can be recognized as a special distance-based technique where the distance of two records is either 1 or 0 [27]. When large databases are linked, this task is computationally very expensive and a strong improvement of its efficiency can be obtained by adopting an appropriate blocking strategy in order to reduce the number of candidate records comparisons [28,29].

Table 1
Matching attributes and related matching functions used for duplicate detection.

Blocking attribute	Match function
Name	Distance function (DL) lower than a threshold value
Surname	Perfect match
Sex	Match
Birth/death date	Overlapping birth/death date intervals

In duplicate detection, we blocked on the attributes reported in Table 1, by adopting the following blocking scheme [30]:

$$\{match, surname\} \wedge \{DL < threshold, name\} \wedge \{match, sex\} \wedge \{overlaps, birth/death\ interval\}$$

Once obtained the block of candidate duplicates, each block element is considered for further comparisons. In particular, several checks are done on the parents demographical data, based on the same rule, to confirm the hypothesis that the candidates refer to the same individual. If the hypothesis is confirmed, one of the two vital records is removed after merging its data with the remaining one. The same kind of data merge is performed also on the data of the individual's relatives.

2.3. Pedigree generation

The pedigree generation algorithm processes a list L of 1st category individuals in order to reconstruct their trios. Additional 1st category individuals (called *substitute* individuals) are generated for the spouses in marriage records. These substitute individuals are added to L so that individuals who are not represented in birth, death or EMR records may be included in the pedigree. Once L is populated, for each individual I_0 in L, the pedigree generation process looks for his/her parents among the individuals in L. A sequence of three steps is necessary to validate the relationship child–parent (Fig. 2).

Step 1 – For each individual I_0, the algorithm starts by searching his/her parent among the demographic data reported in the owner record of each individual in L. If a parental relationship between two individuals is a priori known, the relationship can be manually forced and the procedure is stopped.
Step 2 – For each potential parent found, the algorithm searches for his marriage record. For each marriage record found, the data of the spouse's (husband/wife) parents are compared with the ones of the potential parent. If they match, the algorithm goes to Step 3, otherwise the child–parent relationship is rejected. The availability of the marriage records is mandatory and the child-parent relationship is rejected if no parents marriage records are found. In the Supplemental Figure 1a an example of pedigree is shown. If the marriage act of the parents of Cogo Teresa Maria Luigia (red circle) was eliminated, the tree was interrupted and two fictitious parents were then created from her birth act (Supplemental Figure 1b).
Step 3 – The potential parent spouse's (wife/husband) data are compared with the data of the parent reported in the record of individual I_0. In case of match the child–parent relationship is accepted.

At the end of these steps the pedigree reconstruction process can have one of the following outcomes:

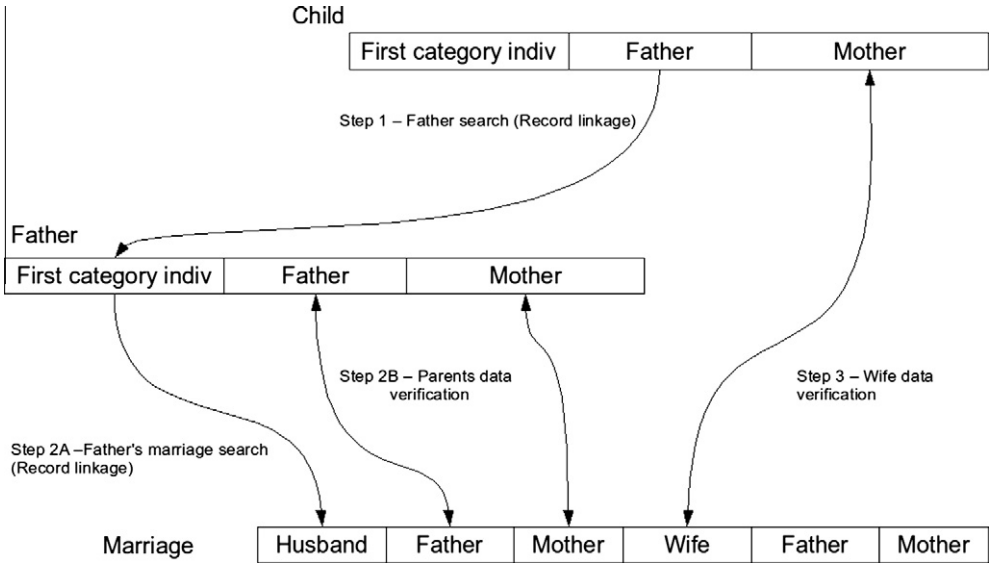

Fig. 2. Scheme of the three steps necessary to validate a child–parent relationship.

(a) parent *forced*: the algorithm uses a priori knowledge about the parenthood;
(b) parent *found*: the parent is found;
(c) *multiple* parents: more than one potential parent is found. This situation generates a multiple relation which data are saved into a report used to detect the possible causes of ambiguity and to manually revise the data;
(d) parent *not found*: if a parent is not found as 1st category individual, a *fictitious* individual is created with demographic data in the son's birth record. This step is necessary to reconstruct the entire family and link brothers. Each fictitious individual generates a trio with parents conventionally set to 0 (corresponding to *unknown*).

This procedure makes the pedigree reconstruction computationally rather precise, but also quite expensive. As for duplicate detection blocking attributes are used for improving the efficiency of the algorithm. In this case the blocking scheme is:

$$\{\text{cluster}, \text{surname}\} \wedge \text{DL} < \text{threshold}, \text{name} \wedge \text{match}, \text{sex} \wedge \text{overlaps}, \text{birth interval}$$

For example, if the father name was "De Coghi" and birth year in the range 1814–1816, a first category individual with surname "de Cuoghi" and a birth year 1815 can be selected since the surname belongs to the same cluster and the birth year is in the range.

At the end of this procedure each child–parent relationship is recorded in a table that will be used to reconstruct the overall pedigree as set of trios. This is saved as a linkage file.

2.4. Reports evaluation

During the last step several files are generated as tools for checking the pedigree completeness and the possible inconsistencies in the data sources. Reports allow speeding up the manual correction process by categorizing missing links so that users can resolve them in a controlled and quicker way. For example, in order to manually link a missing parent it is possible to start looking in the "multiple relations" report that stores all possible parents already found during the last run; this approach may shorten the search time with respect to performing an ex-novo manual search in the acts database. Successful searches are then stored in the database in form of *forced* relations between individuals to be used directly in the next run.

2.5. Implementation

The algorithm has been implemented in Java 1.5 [http://java.sun.com/] and tested on both Unix and Windows platforms. The analysis performed on the Val Borbera dataset described in this paper has been carried on a Linux workstation with 2xIntel(R) Core(TM)2 CPU 6300 @1.86 GHz and 4 GB of RAM using Ubuntu Hardy 8.04 distribution [http://www.ubuntu.com/]. The overall data set has been imported into a database MySQL server version 5.0.45 [http://www.mysql.com/]. The software is structured into several modules, corresponding to the steps described in the previous section, which may run either separately or linked through a complete pipeline. A configuration file is used to setup the run parameters, like the steps to be executed, the threshold for the string matching distance, the minimal and maximal procreative ages, etc. The website of the software reports a detailed description of the configuration file.

3. Results

The software was developed to reconstruct the genealogy of the Val Borbera population [31]. We analyzed the vital records coming from municipal and parish registries, available since 1565 as well as the EMR that contain vital records of the population living in the villages since around 1900 and computerized since 1985. All the 1803 phenotyped individuals enrolled in the Val Borbera project [http://www.valborbera.org/] were included.

The data set contained a total of 94,105 records that represented the input to the pedigree reconstruction process. Table 2 reports for each data source the number of records imported and the number of individuals in the UDB for each category of individuals, as described in Methods, before and after data revision and duplicate merging. 1st category individuals were 66,562. 2nd category included 40,430 individuals. The 3rd category group contained all the remaining individuals reported in each record, namely 250,206 individuals.

The algorithm was designed to reconstruct the trios of all the 1st category individuals (66,562) as well as of the additional substitutes individuals (17,441). The quality and the amount of the

Table 2
Number of records imported from each data source and number of individuals inserted into the UDB for each category before and after duplicate detection.

Data source	Acts	Individuals		
		Imported	Duplicated	After merging
EMR	4637	4637[a]	7[a]	4630[a]
		11,867[c]	15[c]	11,852[c]
Birth registry	64,992	64,992[a]	4561[a]	60,431[a]
		182,212[c]	8938[c]	173,274[c]
Marriage registry	20,793	41,586[b]	1156[b]	40,430[b]
		69,633[c]	8848[c]	60,785[c]
Death registry	3683	3683[a]	2182[a]	1501[a]
		8567[c]	4272[c]	4295[c]
Total	94,105	387,177	29,979	357,198

[a] First category.
[b] Second category.
[c] Third category.

information associated to each 1st category individual were very variable: data often were missing or was approximate. The greatest difference was found between acts dated before or after 1838, when the Napoleon code was introduced in Piedmont and the quality of the municipal and church records greatly improved. This is clearly shown in Supplemental Tables 1a and 1b the were the percentage of records containing name, surname and birth date interval >6 years (see date of birth calculation in Materials and Methods) was calculated for parents in birth acts and for spouses and spouses parents in marriage acts. In most older birth acts the mother's surname was present in only 35% of the acts and the mother date of birth was approximate to >6 years in all, while the same information was present in the majority of the post 1838 acts (Supplemental Table 1a). In the older marriage acts, the fathers' spouses' names and surnames were often missing and the mothers was missing in >99% of the cases. The date of birth of the two spouses was approximate to >6 years in the majority of the older acts (Supplemental Table 1b).

As the quality of the information associated to each individual is a major factor in genealogy reconstruction, not all the 84,003 individuals in our data set were used to reconstruct the trios. As some of the acts are not fully populated in terms of individuals (mainly the father or the mother data are missing), the algorithm tried to find the parents of only 75,994 individuals, as discussed in the next paragraph.

3.1. Method evaluation

After each run we were able to classify the reconstructed trios in the following groups:

- *found*, trios in which both parents were found as 1st category or substitute individuals;
- *fictitious*, trios in which both parents are fictitious individuals. As explained in Section 2.3, fictitious trios are necessary to reconstruct entire families with multiple children;
- *forced*, trios including forced relations manually inserted by archivists. One of the two parents could have been found as 1st category or substitute individual;
- *multiple*, trios with multiple relations for both parents;
- *mixed, all remaining trios*;

The results obtained for the Val Borbera data set are reported in Table 3 and details are given in Supplemental Tables 2a and 2b where the number of trios obtained by any combination of mother and father *found, fictitious, forced and multiple* is reported. Supplemental Table 2a refers to the pedigree derived by exploiting also forced relations, while Supplemental Table 2b shows the results obtained without forced relations. From the data it is clear that most of the trios were reconstructed by the algorithm and the manual intervention was a very minor part of the process: the total number of trios *found* increased only of 4% if forced relations were considered. On the other hand, mixed and multiple trios were reduced by 13% and 18% respectively, indicating that the insertion of forced relations will establish few novel links, but may considerably contribute to pedigree construction by linking smaller pedigrees.

Forced relations were mainly used for completion of pedigrees including distant ancestors and helped in linking smaller pedigrees. The analysis of the trios confirmed that the algorithm performs better when it is applied on good quality data, as expected. If we consider the distribution of the different kind of trios by period and type, we confirmed a large increase of "found" trios and a decrease of mixed and multiple ones after 1838. However, despite the worst quality and limited availability of the data before 1838 and the problems related to the language evolution along centuries, we obtained quite satisfactory results as large numbers of trios were identified also from pre-1838 records (28,848, corresponding to the 44% of the total). Moreover the number of trios over the centuries followed the same distribution as the population size (as measured by the number of births/year) (see Supplementary Figure 2). The number of trios that the algorithm constructed follows the same general distribution per year showing a fast increase around 1850 and a decrease from the beginning of 1900. We can thus conclude that the performance appears quite satisfactory, given the quality of the input data, possibly as the method accepts as true only those trios, which have fulfilled a large number of requirements and cross-checks (see Section 2.3).

Using the algorithm, we were able to reconstruct the pedigree for all the participants to the Val Borbera population project, that included 1803 participants: 88% were clustered in one family of 10,469 persons going back to 13 generations. Finally, the performance of the algorithm was tested by comparison of the kinship matrix determined from the Val Borbera participants' genealogy to the genomic kinship matrix. 1664 DNAs from the Val Borbera cohort were genotyped with an Illumina 370K array. From the genotypes, a genomic kinship matrix can be determined using the GeneAble software [32] which is based on comparison of the 370K SNPs genotyped in 1664 individuals and the construction of a IBD (Identity By Descent) matrix. The Val Borbera genealogy kinship matrix was determined using the KinInbcoef algorithm (KinInbcoef 1.0 (http://www.stat.uchicago.edu/~mcpeek/software/KinInbcoef/index.html). The average genealogy kinship was 4.9×10^{-4}, corresponding to 11 meiosis separating two individuals which may therefore be related through a great grand-father 5/6 generations back. The minimum kinship was 1×10^{-6}. We used the GeneAble software to compare pedigree and genomic kinship matrices, by evaluating the Pearson correlation on all pairwise kinship coefficients for each couple of individuals, namely 1.3×10^{6} couples of genotyped individuals. Around 500,000 people had a kinship >0.

Table 3
Trios and links generated after the pedigree generation.

	Trios kind					Links kind	
	Found	Fictitious	Forced	Multiple	Mixed	Found	Forced
With forced relations	25,009	31,051	2942	1924	14,090	62,065	6862
Without forced relations	24,014	34,472	0	2355	16,173	61,559	0

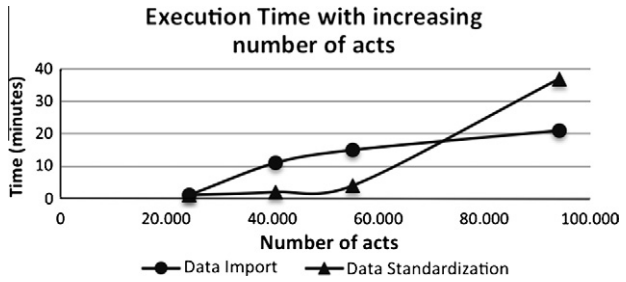

Fig. 3. Execution times of data import and data standardization steps with different dimension input data.

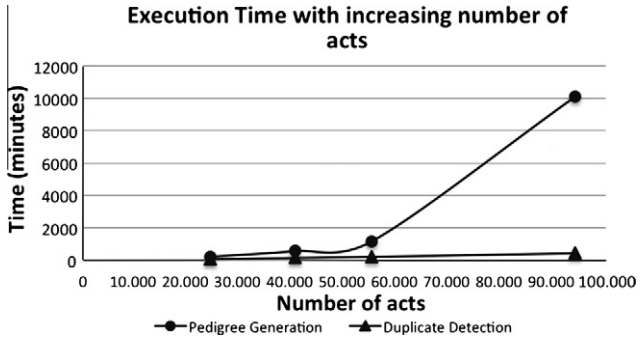

Fig. 4. Execution times of duplicates detection and pedigree generation with different dimension input data.

The correlation describes the degree of relationship (linear dependence) between two variables X and Y, giving a value between +1 and −1. A positive value for the correlation implies a positive association while a negative value implies an inverse association. Correlation between genealogy and genomic kinship was 0.92, this high correlation could be hardly obtained by chance and confirms the quality of the VB genealogy induced so far.

3.2. Performance evaluation

The algorithm scalability was studied by performing four runs with an increasing number of acts coming from one (29,827 documents), two (46,923 documents), three (62,798 documents) and finally all seven villages of the Valley (94,105 documents). The runs have been performed without forced relations. We registered the computational time spent for each separate step: data import, data standardization, duplicates detection and pedigree generation.

Figs. 3 and 4 report the different times spent in the single steps of the processing. As expected most of the time is spent in the steps requiring record linkage operations (pedigree generation and duplicate detection), while the time for the pedigree generation increases very quickly, the time for duplicated detection is almost linear in the number of acts. This is justified by the intrinsic structure of the pedigree generation step that executes, for each individual, many record linkage operations (for each parent both the birth record and the marriage act are searched).

4. Conclusions

In this paper we present an open source software tool designed to perform pedigree reconstruction through record-linkage techniques, starting from the analysis of civil and church registries or any other pedigree information. The tool is made of three modules, which perform different data processing steps: data import, data preparation and cleaning and pedigree building. Each step may be completed independently if the pipeline is properly followed. This allows to repeat some parts of the algorithm when necessary or to exploit only some of the offered functionalities. The data model was designed to be independent from the data sources, so that the algorithm can be applied to heterogeneous information and is completely general and not bounded to the specific case presented in this paper. Moreover, all the linguistic, historical and behavioral aspects of names and surname evolution in the population under study can be easily customized. Finally, a set of methods is available to deal with the name/surname clustering problem, thus supporting the effective reconstruction of a pedigree even dealing with ancient registries. Although the pedigree reconstruction is still semi-automatic, the tool allows to greatly speed up the overall process by managing incremental updates of the pedigree due to new data or to the inclusion of prior knowledge through forced relations.

The tool is best fitted for reconstruction of genealogies of medium size endogamic populations where civil and church registries are available for several centuries and was tested on the reconstruction of the pedigree of the isolated population of the Val Borbera, located in North West Italy. This task was particularly challenging, because of the large and heterogeneous data set related to a long time span: the Val Borbera was a relatively inhabited valley with around 10,000 people living in seven villages in the 19th century. Parish and municipal registries were available since the middle of 1500 and were used for data collection, as well as electronic ones. The strategy implemented allowed to obtain a large number of trios that could be connected to build a large pedigree. A single very large family of more than 100 thousands individuals connected almost 50% of the whole population. The performance of the algorithm was nearly optimal in presence of well-curated records, as shown by the high number of trios generated on the basis of the data collected after 1838 from better organized modern archives. It is interesting to note, however, that the results obtained on the data records collected before 1838, which are often incomplete and imprecise in terms of names/surnames specifications, allowed to map the genealogy of the population under study up to 13 generations. An independent evaluation of the correctness of the algorithm came from the very high correlation (0.92) between genomic and pedigree kinship determined for a genotyped subset of the Val Borbera population. As already mentioned, the tool was designed to be general enough to be used by researchers to effectively support the cumbersome task of pedigree reconstruction and to contribute to the genetic analysis of entire populations deriving from few ancestors, like the isolated population of Val Borbera and that may represent an important tool for the study of genetic variation and of complex disorders.

The software is available for under GNU license at the web site: http://projects.labmedinfo.org/tree/index.html for further use and evaluation.

Acknowledgments

We thank the Val Borbera administrators and the Tortona and Genova archidiocese for continuous support. We are indebted to all the people that helped in data collection, Ilaria Daglio, Federica Lovotti, Elisa Terragno, Francesco Vella, Andrea Repetto, and the archivists Gabriella Parodi and Laura Goggiano. Finally, we thank Angelo Nuzzo for his contribution at the beginning of the tool development.

Funding: The work was supported by the FIRB Project ITALBIONET 'Rete Italiana di Bioinformatica' to R.B., Compagnia San Paolo Torino, Cariplo Foundation, TelethonFoundations and Ministry of Health Progetti Finalizzati 2007 to D.T.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2011.08.004.

References

[1] Altshuler D et al. Genetic mapping in human disease. Science 2008;322(5903):881–8.
[2] Freimer N, Sabatti C. The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. Nat Genet 2004;36(10):1045–51.
[3] Varilo T, Peltonen L. Isolates and their potential use in complex gene mapping efforts. Current Opin Genet Dev 2004;14(3):316–23.
[4] Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet 2009;41(1):35–46. Epub 2008 Dec 7.
[5] Jakkula E, Leppä V, Sulonen AM, Varilo T, Kallio S, Kemppinen A, et al. Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in STAT3 gene. Am J Hum Genet 2010;86(2):285–91.
[6] Stacey SN, Sulem P, Zanon C, Gudjonsson SA, Thorleifsson G, Helgason A, et al. Ancestry-shift refinement mapping of the C6orf97-ESR1 breast cancer susceptibility locus. PLoS Genet 2010;6(7):e1001029.
[7] Agarwala R et al. Software for constructing and verifying pedigrees within large genealogies and an application to the old order Amish of Lancaster County. Genome Res 1998;8:211–21.
[8] Henneman P, Aulchenko YS, Frants RR, van Dijk KW, Oostra BA, van Duijn CM. 2008 Prevalence and heritability of the metabolic syndrome and its individual components in a Dutch isolate: the Erasmus Rucphen Family study. J Med Genet 2008;45(9):572–7.
[9] Gulcher J, Stefansson K. Population genomics: laying the groundwork for genetic disease modeling and targeting. Clin Chem Lab Med 1998;36(8):523–7.
[10] Riester Markus, Stadlerand Peter F, Klemm1 Konstantin. FRANz: reconstruction of wild multi-generation pedigrees. Bioinformatics 2009;25(16):2134–9.
[11] Konovalov DA, Manning C, Henshaw MT. KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. Mol Ecol Notes 2004;4:779–82.
[12] Ashley MV, Caballero IC, Chaovalitwongse W, Dasgupta B, Govindan P, Sheikh SI, et al. Kinalyzer a computer program for reconstructing sibling groups. Mol Ecol Resour 2009;9(4):1127–31.
[13] Mancosu G, Cosso M, Marras F, Borlino CC, Ledda G, Manias T, et al. Browsing isolated population data. BMC Bioinform 2005;6(Suppl. 4):S17.
[14] Alter George, Mandemakers Kees, Gutmann Myron P. Defining and distributing longitudinal historical data in a general way through an intermediate structure. Social Res 2009;59:78–114.
[15] Wang JR, Madnick SE. The inter-database instance identification problem in integrating autonomous systems. In: Proceedings fifth international conference on data engineering; 1989. p. 46–55.
[16] Damerau F. A technique for computer detection and correction of spelling errors. Commun ACM 1964;7(3):171–6.
[17] Levenshtein V. Binary codes capable of correcting spurious insertions and deletions of ones. Prob Inf Transm 1965;1:8–17.
[18] Levenshtein V. Binary codes capable of correcting deletions, insertions and reversals. Sov Phys Dokl 1966;10(8):707–10. Original in Russian in Dokl. Akad. Nauk SSSR 1965; 163(4): 845–8. English translation in Soviet Physics Doklady 10(8): 707–10.
[19] Newcombe Howard B, Kennedy James M, Axford SJ, James AP. Automatic linkage of vital records. Science 1959;130(3381):954–9.
[20] Newcombe HB. Record linking: the design of efficient systems for linking records into individual and family histories. Am J Human Genet 1967;19:335–59.
[21] Fellegi IP, Sunter AB. A theory of record linkage. J Am Stat Assoc 1969;64:1183, 1210.
[22] Adams Melissa M, Wilson Hoyt G, Casto Dale L, Berg Cynthia J, McDermott Jeanne M, Gaudino James A, et al. Constructing reproductive histories by linking vital records. Am J Epidemiol 1997;145:339–48.
[23] Herrchen B, Gould Jeffrey B, Nesbitt Thomas S. Vital statistics linked birth/infant death and hospital discharge record linkage for epidemiological studies. Comput Biomed Res 1997;30:290–305.
[24] Christen P. Automatic record linkage using seeded nearest neighbour and support vector machine classification. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, August 24–27, 2008, Las Vegas, Nevada, USA.
[25] Bilenko M, Mooney RJ, Cohen WW, Ravikumar P, Fienberg SE. Adaptive name matching in information integration. IEEE Intell Syst 2003;18(5):16–23.
[26] Hernandez MA, Stolfo SJ. Real-world data is dirty: data cleansing and the merge/purge problem. Data Min Knowledge Discov 1998;2:9–37.
[27] Elmagarmid AK et al. Duplicate record detection: a survey. IEEE Trans Knowledge Data Eng 2007;19:1–16.
[28] Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J Am Stat Assoc 1989;84(406):414–20.
[29] Baxter R et al. A comparison of fast blocking methods for record linkage. In: Proceedings of KDD-2003 workshop on data cleaning, record linkage, and object consolidation; 2003.
[30] Michelson M, Knoblock CA. Learning blocking schemes for record linkage. In: Proceedings of the 21st national conference on Artificial intelligence, July 16–20, 2006, Boston, Massachusetts. p. 440–5.
[31] Traglia M et al. Heritability and demographic analyses in the large isolated population of val borbera suggest advantages in mapping complex traits genes. PLoS One 2009;4(10).
[32] Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics 2007;23:1294–6.